Project Proposal: Adversarial Training for Robust Text Generation

Sebastien Baur
sebastienbaur@g.harvard.edu

Bernd Huber
bhb@seas.harvard.edu

Srivatsan Srinivasan
srivatsansrinivasan@g.harvard.edu

March 23, 2018

1 Problem Description

1.1 Definition

The project aims to construct robust generative models using adversarial training(GAN variants). We particularly intend to tackle the issue of semantic coherence in the longer sentences generated by proposing alterations to the minimax objectives, model architectures and training procedures and compare their efficacy quantitatively and qualitiatively with our chosen baseline. Using ARAE-GAN (Kim et al. (2017)) as our baseline, we suggest the following enhancements - replace Lipschitz constraint on the critic with penalizing the critic's gradient norm as suggested by (Gulrajani et al. (2017)) and evaluate progressive training of GANs - training on smaller sentences and larger ones subsequently to ensure gradual feature transfer in latent space (Karras et al. (2018), Press et al. (2017)). Based on the progress and deficiencies of initial steps and time constraints, we also plan to experiment with single generator and multiple discriminators(focusing on different temporal lengths of sentences), multiple generators(attempting to prevent mode collapse) with single discriminators and alternate training architectures such as LeakyGAN(Guo et al. (2017)), RankGAN(Lin et al. (2017)), stackGAN(Zhang et al. (2016)) etc.

1.2 Introduction and Relevance

While NNLM, LSTM etc. have been successful in language modeling task, adversarially trained generative models arise three-pronged interest in language generation

1. They have proven extremely successful in image generation in synthesizing high quality diverse images.
2. Due to learning a latent distribution in the continuous space, we get far superior expressiveness and variability in the generated text(via the variance learned in the latent space) and create additional scope for conditioning, transfer etc.
3. They are less susceptible to exposure bias which plague common language and seq-to-seq models.

On the other hand, creating such generative models on text is extremely difficult because the latent variables are learned in the continuous space while the text is discrete. To get around this issue, (Kim et al. (2017)), presents a model that jointly trains a discrete space autoencoder(RNN) and a continuous space generation function using GANs by constraining similarity between distributions. While this model performs admirably in intended tasks such as style transfer, mode collapse and lack of generative robustness(based on the groups' conversations with Yoon) still prevail. Meanwhile, there has been a recent flurry of alternate adversarial architectures and critic constraints(both in the images and text universe) that have been proposed, few of which we intend to explore through this work. As stated earlier, a key aspect we intend to target is an architecture that is robust enough to produce semantically coherent longer sentences which are also diverse.

2 Plan of Action

We propose to incrementally unfold our project with the following steps and these alternatives are conditional on the results and issues that occur in the previous step. Many of these approaches have been successful in the space of images and we would like to experiment with their efficacy on natural language generation.

1. Recreate baselines - ARAE-GAN (Kim et al. (2017)), SeqGAN (Yu et al. (2017))
2. WGAN improvements - Replace Lipschitz constraint with penalty on critic gradient norm proposed by (Gulrajani et al. (2017)) or spectral norm regularization of discriminator proposed by (Miyato et al. (2018))
3. Progressive/Curriculum Training on GANs (Karras et al. (2018), Press et al. (2017)) - Sequentially train GANs on smaller and larger sentences while transferring latent features in the process.
4. Experiment newer architectures - multiple generators and discriminators, LeakyGAN(Guo et al. (2017)), RankGAN(Lin et al. (2017)), stackGAN(Zhang et al. (2016)) and our own similar variants to discriminator and/or generator objectives.

3 Dataset

We will use the existing corpus: Stanford Natural Language Inference: The SNLI corpus (version 1.0) is a collection of 570k human-written English sentence pairs manually labeled for balanced classification with the labels entailment, contradiction, and neutral, supporting the task of natural language inference (NLI), also known as recognizing textual entailment (RTE). We aim for it to serve both as a benchmark for evaluating representational systems for text, especially including those induced by representation learning methods, as well as a resource for developing NLP models of any kind. Besides, there are standard sentence datasets such as Yelp and Yahoo datasets which provide viable alternatives incase we run into any issues with SNLI data.

4 Evaluation Methods

We will derive lot of methods similar to the ones expressed in ARAE-GAN (Kim et al. (2017)). We will use the reverse perplexity - which is calculated by training a language model on the generated data and measuring perplexity on real hold-out data and also the normal perplexity to measure the fluency of the generated text. Corpus level BLEU score is another alternative quantitative metric that we propose to undertake. We will also demonstrate qualitative evaluations of random set of samples generated by different architectures and try to reason the samples generated with respect to the architectures of GANs. The conditional GANs will additionally be evaluated by testing the generated sentences using a pre-trained classifier. We will also evaluate whether the model is memorizing sentences implicitly ignoring the latent space entirely, by comparing the synthetic sentences with real corpus for similarity and also check if modifying latent variables indeed provides variability in our sentences. In terms of models, our experiments have already been outlined in plan of action and each successive experiment will be a sequential one step-addition to the most successful variant of the previous experiment.

5 Related Work

Wasserstein GANs have been an extremely successful architecture in image generation. There have been several variants of the W-GAN that primarily alter the discriminator gradient constraints from the original work and have proven to perform better than the baselines.(Miyato et al. (2018), Gulrajani et al. (2017)). Another successful GAN training method from the world of images has been progressive GANs(Karras et al. (2018)) where an image GAN is progressively trained from low-res images to really high res-images thus ensuring gradual learning of the highly complex latent space, which empirically proves better and faster learning capabilities.

Despite the success stories and large focus, the discriminator in GANs is still mostly limited to a binary classification task. This limited learning capacity represents a major drawback for tasks where the generated data is highly structured, such as natural language. There are multiple ways that researchers tried to circumvent this issue. In the RankGAN paper, the discriminator task is turned into a ranking task by letting the discriminator to rank relative goodness of outputs, rather than binary classification (Lin et al., 2017).

The limitation of discriminators is especially challenging during the generation of longer sentences, since the discriminator will only give a signal at the very end of a sentence. To circumvent this problem, leaky discriminator structures have been tested by Guo et al. (2017). Specifically, the features extracted by the discriminator are encoded by latent(context) variable that provides feedback to the generator at any given time. The results of this research show that generated sentences then become more coherent over longer sequences.

Generative language modeling has also been associated with policy optimization and imitation learning literature. While the classic discriminator architectures provide classification at the end of the episode(sentence), having the ability to run policy update throughout the input sequence offers more granular information to the generator making it more robust. This idea has been explored through policy gradient methods in the work of SeqGan((Yu et al., 2017)). Using policy optimization methods has been further tried in imitation learning, where learning occurs through demonstration, rather than feedback loops Ho and Ermon (2016).

6 Proposed Milestones

1. March 30 - Literature Survey and Baseline Recreation
2. April 11 - First set of alternatives(WGAN, curriculum training etc.) trial and comparison with baselines
3. April 12 - In-class presentation sign-up
4. Rest of April - Incorporation of class feedback and second set of alternatives(optional - try out other architectures)
5. Early May - Final presentation
6. May 8 - Report completion

References

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. (2017). Improved training of wasserstein gans. *arXiv preprint arXiv:1706.01399*.

Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Yong, and Wang, J. (2017). Long text generation via adversarial training with leaked information. *arXiv preprint arXiv:1709.08624*.

Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kim, Y., Zhang, K., Rush, A. M., LeCun, Y., et al. (2017). Adversarially regularized autoencoders for generating discrete structures. *arXiv preprint arXiv:1706.04223*.

Lin, K., Li, D., He, X., Zhang, Z., and Sun, M.-T. (2017). Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pages 3155–3165.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks.

Press, O., Bar, A., Berant, J., and Wolf, L. (2017). Language generation with recurrent generative adversarial networks without pre-training. *arXiv preprint arXiv:1706.01399*.

Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858.

Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. (2016). Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*.